2023

Mitigating Anomalous Electricity Consumption in Smart Cities Using an AI-Based Stacked-Generalization Technique

Arshid Ali
COMSATS Institute of Information Technology, Pakistan

Laiq Khan
COMSATS Institute of Information Technology, Pakistan

Nadeem Javaid
COMSATS Institute of Information Technology, Pakistan

Safdar Hussain Bouk
Old Dominion University

Abdulaziz Aldegheishem
King Saud University

See next page for additional authors

Original Publication Citation

Ali, A., Khan, L., Javaid, N., Bouk, S. H., Aldegheishem, A., & Alrajeh, N. (2023). Mitigating anomalous electricity consumption in smart cities using an AI-based stacked-generalization technique. *IET Renewable Power Generation*. Advance online publication. https://doi.org/10.1049/rpg2.12785

Received: 6 March 2023 | Revised: 17 May 2023 | Accepted: 19 June 2023

IET Renewable Power Generation

DOI: 10.1049/rpg2.12785

The Institution of Engineering and Technology WILEY

ORIGINAL RESEARCH

Mitigating anomalous electricity consumption in smart cities using an AI-based stacked-generalization technique

Arshid Ali[1] | Laiq Khan[1] | Nadeem Javaid[2] | Safdar Hussain Bouk[3] | Abdulaziz Aldegheishem[4] | Nabil Alrajeh[5]

[1]Department of Electrical and Computer Engineering, COMSATS University Islamabad, Islamabad, Pakistan

[2]Department of Computer Science, COMSATS University Islamabad, Islamabad, Pakistan

[3]Department of Computer Science, Old Dominion University, Norfolk, Virginia, USA

[4]Department of Urban Planning, College of Architecture and Planning, King Saud University, Riyadh, Saudi Arabia

[5]Department of Biomedical Technology, College of Applied Medical Sciences, King Saud University, Riyadh, Saudi Arabia

Correspondence
Nadeem Javaid, Department of Computer Science, COMSATS University Islamabad, Islamabad 44000, Pakistan.
Email: nadeemjavaidqau@gmail.com

Funding information
King Saud University, Grant/Award Number: RSP2023R295

Abstract

Energy management and efficient asset utilization play an important role in the economic development of a country. The electricity produced at the power station faces two types of losses from the generation point to the end user. These losses are technical losses (TL) and non-technical losses (NTL). TLs occurs due to the use of inefficient equipment. While NTLs occur due to the anomalous consumption of electricity by the customers, which happens in many ways; energy theft being one of them. Energy theft majorly happens to cut down on the electricity bills. These losses in the smart grid (SG) are the main issue in maintaining grid stability and cause revenue loss to the utility. The automatic metering infrastructure (AMI) system has reduced grid instability but it has opened up new ways for NTLs in the form of different cyber-physical theft attacks (CPTA). Machine learning (ML) techniques can be used to detect and minimize CPTA. However, they have certain limitations and cannot capture the energy consumption patterns (ECPs) of all the users, which decreases the performance of ML techniques in detecting malicious users. In this paper, we propose a novel ML-based stacked generalization method for the cyber-physical theft issue in the smart grid. The original data obtained from the grid is preprocessed to improve model training and processing. This includes NaN-imputation, normalization, outliers' capping, support vector machine-synthetic minority oversampling technique (SVM-SMOTE) balancing, and principal component analysis (PCA) based data reduction techniques. The pre-processed dataset is provided to the ML models light gradient boosting (LGB), extra trees (ET), extreme gradient boosting (XGBoost), and random forest (RF), to accurately capture all consumers' overall ECP. The predictions from these base models are fed to a meta-classifier multi-layer perceptron (MLP). The MLP combines the learning capability of all the base models and gives an improved final prediction. The proposed structure is implemented and verified on the publicly available real-time large dataset of the State Grid Corporation of China (SGCC). The proposed model outperformed the individual base classifiers and the existing research in terms of CPTA detection with false positive rate (FPR), false negative rate (FNR), F1-score, and accuracy values of 0.72%, 2.05%, 97.6%, and 97.69%, respectively.

1 | INTRODUCTION

The successful integration of renewable energy sources into the electricity network transformed the power grid from a centralized and dull energy system to a decentralized and intelligent system. This distributed power system makes the grid more efficient due to efficient infrastructure utilization. The recent technological development and new strategies followed by the utilities make the grids more flexible for energy resource accumulation. Therefore, more intermittent energy resources can be used for electricity generation and are added to the power system without disturbing the grid stability. According to US Energy Information Administration (EIA), an increase in electricity generation from renewable sources is more than 20%

[1]. In addition to the need for improvement in the amount of electricity generation by adding more resources to the electric grid, power management and efficient energy resource utilization also play a useful role in the socioeconomic development of a country. It is because of the high cost of electricity production and limited availability of energy resources. To cut down on the electricity bills and have uninterrupted flow of electricity, the consumers are directed towards anomalous electricity consumption. There are many ways of consuming electricity in an anomalous manner; electricity theft is being one of them. To deal with such electricity consumption, efficient power management and cost reduction are considered. Power management and cost reduction are possible in two ways.

1. Generate and transmit electricity from those resources that have minimum expense per unit.
2. Rated revenue pay-back of consumed electricity to utility in the form of the electricity billing system.

The reduction in unit per cost of electricity can be addressed by moving towards low-cost, low-emission renewable sources with more energy-efficient devices. While the revenue pay-back system of utility faces issues due to electric power loss (EPL). The difference between the energy generated at the generation end and the energy delivered to the consumers is known as electricity loss. The electricity losses are classified into two categories [2].

1. Technical losses or system losses (TLs)
2. Non-technical losses (NTLs)

TLs are the total EPL in the power system, from the network injection point to the consumption point. These occur due to energy dissipation in transmission lines, distribution lines, and transformer cores. This problem can be solved by using good quality and highly efficient equipment instead of old electrical infrastructure, but this process requires a huge economic investment and it is time consuming. NTL may be due to some kind of abnormality or changes induced by electricity consumers (EC) in the electricity network like installation errors, billing errors, faulty meters, or meter by-passing. This creates system disturbance and low power load management for utility companies. In addition, NTLs or electricity theft (ET) not only cause significant economic loss but also affect the normal operations of the power system by creating power fluctuations and disturbing grid stability [3]. According to Northeast Group, the NTL-based worldwide revenue losses were about $96 billion in 2017 [4]. While in 2014, these losses were about $58.7 billion in the world with India facing 16.2 billion USD, Brazil facing 10.5 billion USD, Pakistan facing 0.89 billion USD, and Russia facing 5.1 billion USD loss [5, 6], which shows a high increase in loss during the last few years.

The recent technological development, advanced metering infrastructure (AMI) system, and especially smart grids make electricity management, monitoring, and NTL reduction possible. The SG is an intelligent electricity system that permits a two-way flow of electricity and information by using an intelligent monitoring system. It integrates the AMI system to control and monitor the energy usage of consumers and utility in the electricity network [7, 8]. This system works in real-time by first collecting the user's electricity consumption (EC) information and then transferring it to the utility using communication channels for billing, grid security, loss reduction, and other purposes. The collection of EC in real-time makes the SG capable to detect the losses in electricity networks. The two main types of information required about energy loss are given below.

1. "How to locate the theft source?"
2. "How much electricity is stolen?"

The NTL-based losses are mainly experienced by the illegal EC by the users, which also disturbs the system operation, incurs additional losses, damages the system components, and affects the grid security and stability. Many countries have characterized electricity theft as a special kind of crime [9]. To reduce NTLs, utility companies must follow the necessary steps to identify the theft and abnormal behavior of energy usage. However, the conventional methods require a large number of technicians to perform the on-the-spot checkup of the consumption meters. The manual energy consumption reading process lacks organized time and labor schedules. Due to this, an insignificant amount of energy theft is detected, which results in a less revenue pay-back [10].

The recent rapid improvements in ML methods show increased interest in the ideas of models analyzing the load information, and meter tampering as early as possible. The ML theft detection techniques work to detect the deviation of energy statistical patterns from normal behavior. In modern research, the use of ML techniques provides a new solution for utility companies for detecting anomalous EC. These modern techniques make it possible to automate and improve detection accuracy by accurately identifying malicious patterns [11]. Thus, an ML classifier with high accuracy is needed to help the existing techniques in dealing with large theft detection tasks. To overcome the electricity theft issue, many data-driven methods have been used in recent years. These methods are divided into three categories, namely state-based, game theory-based and artificial intelligence (AI)-based [12].

1. State-based methods use specific kinds of devices or designs for metering and theft detection purposes. For example, a special ammeter checks the electricity difference between the local and remote ends for fraud detection purposes. State-based estimation works only at the substation level and not at the end-user level. This type of installation for electricity theft detection (ETD) requires extra monitoring devices, which are difficult to install in the existing distribution systems.
2. Game-theory-based methods use the comparative behavior of pricing competition and product releases like games between anomalous users and electric companies. The main goal of these methods is to find an equilibrium state for the game. This type of model is easy to install but it is

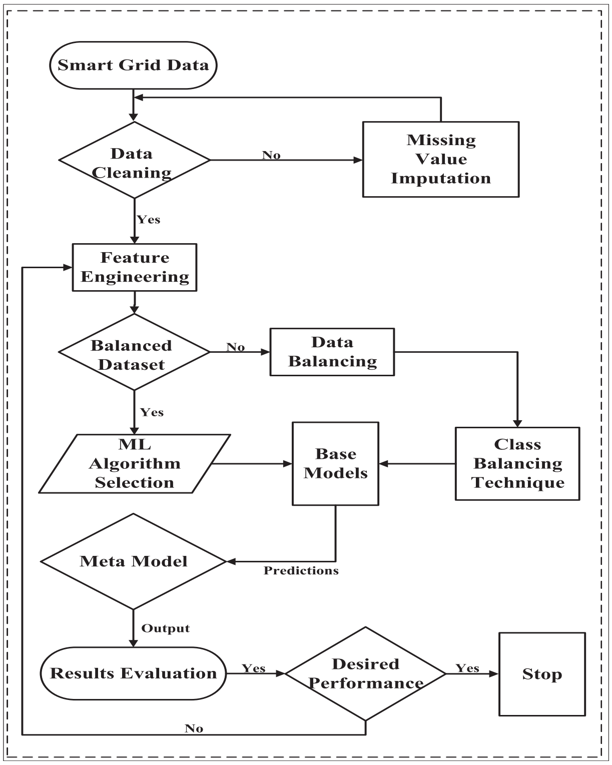

FIGURE 1 Proposed model's flowchart.

hard to find specific mathematical modeling for it, which relates the actual behavior of the end user with the utility company.

3. AI is adopted in almost all worldwide fields including business, security, sales, banking, and many more. The expansion and advancements of smart sensors in SG generate a massive amount of data [13]. This data requires a scalable technique for efficient utilization. The recent advancements in ML and DL in anomaly detection pave a way for energy security in SG [14]. These ML-based models can be used to address the NTL issue in the SG. In the present AMI system, the AI techniques can be used to draw and compare the load profile and the energy consumption pattern of end-users to classify legal and illegal electricity users.

This research aims to present an accurate theft and normal users' classification model using the state grid corporation of China (SGCC) dataset. In this work, we will use some pre-processing steps such as data cleaning and data normalization as shown in Figure 1.

In the preprocessing step, the data is checked for missing values which are NaNs. If the NaN exists, then it is imputed using mean imputation techniques and then the data is normalized. In the following step, the features are reduced which aims to remove the irrelevant features, as th may lacks sufficient information [15]. Thus, the data becomes more useful and it also helps to reduce the time complexity issue and also helps to better express the given data [16]. These steps are followed by data balancing, which aims to overcome the imbalance class issue and improve classification performance using state-of-the-art class balancing techniques. Afterwards, a Stacked (takes the output of ML based models and uses it as an input to the meta-model for output prediction) machine and deep learning (DL) generalization technique is used for improved classification accuracy. Finally, the output results are evaluated, if the desired performance are obtained then the process is stopped.

1.1 | Contributions

It has been observed from the literature that most of the present research works use different intelligent ML methods to detect the NTLs' behavior in the time-series data of SGs. However, the current research still has less accuracy and a research gap in NTLs' behavior detection. The current theft detection issues are tackled in the form of the following contributions.

1. The data obtained from the smart meter has data of both normal and theft users where the number of abnormal users is less than the normal electricity users. Many research works use classification models on the data obtained from the smart meters without considering the issue of class imbalance. The class imbalance biases the ML model towards the majority class and the model classifies theft as a normal user. This class's imbalanced data needs a proper balancing technique to overcome the bianess issue. The SVM-SMOTE class balancing technique help in better addressing the class imbalance issue.
2. The second problem addressed in this study is high dimensionality in the time-series dataset. The high dimensionality causes time complexity issues and reduces output classification performance. This issue is reduced through a proper feature-reduction technique. The PCA feature reduction technique has been used in this paper for dimensionality reduction purposes.
3. Third, many researchers emphasize output results compared with the original labels of the testing set and do not focus on the detection level of abnormal electricity users. The results' comparison in the form of accuracy is not a proper metric. It may result in a set of theft users inspected as normal users, which should be reduced. In the confusion matrix, when the abnormal consumers are predicted as normal consumers, it is considered as a false negative. This false negative issue is addressed in this research to reduce revenue loss.
4. In ML techniques, some normal users are predicted as malicious, which increases the on-spot inspection cost. The fourth contribution in this paper addresses the issue in the form of a maximum reduction in false positive rate.

1.2 | Layout of paper

The paper is divided into eight sections. Section 1 shows the introduction, Section 2 is for related work, Section 3 presents the proposed model's description and the classification

TABLE 1 List of abbreviations and acronyms.

Abbreviations	Full form
AMI	Advanced metering infrastructure
CPTA	Cyber-physical theft attack
PCA	Principal component analysis
SVM	Support vector machine
LGB	Light gradient boosting
ET	Extremely randomized tree
XGBoost	Extreme gradient boosting
MLP	Multi-layer perceptron
RF	Random forest
TNR	True negative rate
FNR	False negative rate
EPL	Electric power loss
NTL	Non-technical loss
TL	Technical loss
SG	Smart grid
ECP	Energy consumption pattern
TPR	True positive rate
FPR	False positive rate
AUC	Area under the curve
NaN	Not a number
Symbols	**Description**
σ	Standard deviation
Z	Z score
μ	Mean value
W	Weight matrix
$x_{m,n}$	Daily energy consumption
N	Number of features
λ	Eigen values
c	Number of classes
b	Bias
x_i	Input data-point
p_i	Probability of outcome of a class

parameters are explained in Section 4. Section 5 explains the simulation setup for the implementation. In Sections 6 and 7, the result discussion and conclusion are provided, respectively. The future work is discussed in Section 8. Most importantly, the list of abbreviations and acronyms, used by the author, is given in Table 1.

2 | LITERATURE REVIEW

This section discusses the existing work done to reduce the NTLs in the SG while considering four main categories of NTL detection methods, including hardware-based, state-based, game theory-based, and AI-based techniques [17]. The methodological structure begins by first having an overview of the classical approach and then moving towards recent artificial intelligence (AI) based techniques. The AI-based ETD steps are then studied, in recent research, focusing on pre-processing, balancing, feature engineering, algorithm modeling, and how these techniques help in the ETD process. NTL has been the major issue in achieving grid stability and causing revenue losses to the utility for more than two decades. Researchers have addressed the issue of NTL reduction in power systems using state-of-the-art techniques available at that time. For example, Pasdar et al. [18] proposed a smart metering system with a high-speed signal to detect malicious activity in the network. The system uses power line communication (PLC) to communicate the customers' energy meter with the utility. In this method, a lossless high-frequency signal, with known line impedances, is transceived through PLC. The software at the utility compares the signals of end users and detects the theft location. A similar case with little modifications in electricity consumption observability is proposed in [19]. The proposed method work using smart energy meters with a specialized display system for both utility and end user. Using the display system, the end users can analyze their own consumption while at the same time utility also monitor and check the consumption behavior. In this way, the power quality and grid stability are maintained. A smart meter with a single chip-based checkup system is implemented in [20] for ETD purposes. The chip uses the standard measurement as the base and then predicts the malicious behavior by comparing it with real-time consumption. The same hardware-based detection methods are proposed in [21–24] with wireless, especially GSM-based monitoring systems. A sensor network with a cloud-based module monitor, circuit breaker, and real-time electricity pulse observer are used to compare and monitor the input and output electricity to the energy meter. The network then uses some form of switch, or buzzer to power off the line and inform the utility about illegal activity. A similar hardware model is presented in [25] based on state-based estimation. The model uses PLC and supervisory control and data acquisition (SCADA) to check the state of connected devices. The PLC is used for communication purposes while SCADA uses internet protocol (IP) services and distributed network protocol 3 (DNP3) for exact device identification and system interoperability, respectively. The acquired data through the state controller module (SCM) is then compared with the standard data produced by the load system (LS) to identify malicious attacks between two connected grids/substations. The above mentioned techniques used specialized sensors, hardware, and online monitoring unit. Besides, the techniques may have hardware failure issues, and can only measure and detect physical theft attacks with no capability of cyberattack detection. Therefore, the authors in [26] proposed a measurement-based approach for NTL reduction. The paper follows the use of an energy monitoring unit on the secondary side of the distribution transformer. The unit takes total electricity consumption measurement and sends the information to the utility of the particular group. The measurement is compared by applying a statistical approach to identity theft among the given group of consumers. Based on the findings of the above hardware-based approaches, the techniques help in NTL reduction. But the system may suffer from a high cost of investment in hardware and unreliability problems. Moving towards its advanced version like AMI, the authors in [27] suggested a multi-source information fusion (MSIF) technique using AMI data for more accurate detection purposes. The data collected from electricity consumers can't be easily classified as malignant or benign based on a single alert. Therefore, a combination of alerts from several malicious users is presented for accurate results. The authors also show the pros and cons of using supervised and unsupervised techniques for output performance. The basic function of the system is that data collected from the AMI system have information about the appliance consumption of that particular customer and also the meter reading. So if a particular device is observed ON and the meter shows zero consumption, then that user is a theft. However, the complete information about customer consumption also causes privacy issues. Further research and advancement to the hardware-based NTL detection technique in the form of AMI, data-driven based techniques including a game theory, and heuristic algorithm-based techniques are applied, such as in [28, 29], to increase the detection accuracy. The main purpose of the techniques is to model a game between electric utility and electricity user. The system is designed to find an equilibrium between the two and a threshold value is assigned. A probabilistic technique is then applied to classify obtained data as honest or ET users. The game theory and heuristic algorithm take sufficient time to deal with big data due to their stochastic nature. These techniques are inaccurate, biased, and can't reach an optimal value on a large dataset. The load flow method based on the AMI dataset is implemented in [30]. The authors addressed the ETD issue in the SG using the real-time electricity consumption pattern (ECP) of consumers. The main problem with power flow analysis techniques, namely Gauss Siedal, Newton Raphson, and Fast Decoupled methods, is that they have low convergence rates, large memory requirements, and time complexity issues in reaching an optimal point. The proposed system addressed this issue by using modified linear regression to capture the electricity theft and normal patterns. This resulted in an increase in the speed of power flow model simulation and its adoption in large power systems. Due to the new technological development in SG, especially AMI systems with real-time monitoring, a large-scale ECP is collected in the form of big data. The newly emerging field of data science (DS) techniques has almost replaced the traditional NTL detection techniques because of the low cost, easy implementation, and high ETD rate. The big data obtained from the utility can be given to DS techniques for easy and efficient analysis. The DS-based machine learning (ML) and deep learning (DL) algorithms have the capability of NTL detection and revenue loss reduction. For example, the authors in [31] proposed a hybrid ML model. Decision tree (DT) and support vector machine (SVM) are used such that extra features are extracted from the original dataset and then fed to the SVM with given features. SVM is used for the final prediction. In addition, data pre-processing is done with missing data imputation and normalization steps. The experiment is done on a dataset collected from various homes in the USA. A non-linear radial basis function (RBF) kernel is selected for SVM to improve the output results. The final accuracy and false positive rate (FPR) obtained were 92.5% and 5.12%, respectively. A similar ensemble approach is also used by Cvitic *et al.* for classification purposes in smart homes [32]. Zhongzong and He in [33] implemented an extreme gradient boosting (XGB) classifier for ETD. The method considers the Irish (Ireland) dataset without using proper data pre-processing, and data balancing techniques. The data reduction is done in the way that six artificial theft attacks were generated from the original dataset and then the model training/testing is performed with that data. The final output obtained is compared with SVM-based classification. The results showed that XGB outperforms SVM in terms of precision, recall, and FPR. A novel idea of gradient boosting classifier (GBC) based theft detection method is proposed in [34]. The research mainly focuses on feature engineering and hyper-parameter tuning steps for improvement in detection rate and FPR, and reduction in processing time. The feature extraction is done using a combination of synthetic feature generation and weighted feature importance (WFI) techniques. The final results showed that GBC outperforms, categorical boosting (CB), light gradient boosting (LGB), and XGB, in terms of FPR and execution time. The authors in [8] proposed a supervised ML technique for all kinds of anomaly detection in SG. For this purpose, Endesa (Spain) dataset is considered which is collected from almost 57000 field inspections of different consumers. The feature extraction is done using ECP, distance, density, and electrical magnitude-based measurement. The Endesa dataset also has important information on geographical, seasonal, and smart meter properties. The final XGB model shows better results with an AUC of 91%. Prem et al. [35] worked on cyber-physical attack detection using an isolation forest classifier (IFC). The isolation forest is used to detect the change in the pattern of the consumers. The main purpose of theft is to decrease the meter reading from actual values, which changes the energy consumption pattern (ECP) of that particular user. Data reduction is done using PCA. The IFC is trained at varying load and voltage generation in order to capture the exact picture from all possible ECP of consumers. The hyper-parameter tuning is done and the model is tested for different grid/bus systems. The results obtained show 98.7% recall in terms of anomaly detection with the IEEE 3-bus system. Leloko et al. [5] tried to differentiate theft consumers from honest consumers using the SGCC dataset. The overall method used data pre-processing, data balancing, and feature reduction. Hyperparameter tuning is done using a Bayesian optimizer. The model was individually implemented on both time domain features and frequency domain features for accurate training. Feature selection from both the time and frequency domains proves useful. The final deep neural network implemented showed outstanding performance of 97% and 91.8% with an area under the curve (AUC) and accuracy, respectively. In [36], Paria et al. presented a solution for ETD purposes while focusing on the ECP of consumers. The ECP of theft and honest users are not the same, in fact, the theft pattern has more fluctuations. Therefore, the area with a high probability of malicious activities, in terms of electricity consumption, are installed with distribution transformer meters (DTM). Using this transformer meter, both types of consumers are identified. 5000 real consumers' data is analyzed in this work. The data preprocessing, balancing, and feature reduction are all overcome by generating six synthetic attacks. SVM algorithm with the combination of noting different types of ECP is obtained using DTM. The final experimental result showed a 93% detection rate and 11% FPR. Similar to the above ECP-based NTL detection, an optimized convolutional neural network and gated recurrent unit (CNN-GRU) method is studied in [37]. Real-time data of 10,000 consumers is analyzed for ETD purposes. The data preprocessing is done to impute missing values. Synthetic minority over-sampling (SMOTE) is used for class balancing. A manta ray foraging optimization (MRFO) is combined with CNN-GRU for result improvement. The final model implemented showed a 91.1% accuracy which was greater than SVM, logistic regression, and CNN-GRU. The same data-driven approach is applied in [38] for NTL reduction in the SG. The authors worked on real-time data of 2,271 consumers collected from the Honduras distribution system. The smoothing spline function (SSF) is used for outlier handling. For feature reduction purposes, a new discrete wavelet packet transform is implemented. The class imbalance issue is addressed using the random under-sampling (RUS) technique. In the last step, an ML-based RUS with Adaboost technique is applied for classification purposes. Adaboost performed better with an accuracy of 94.35% when compared with Linear-SVM, Non-Linear SVM, and artificial neural network (ANN). Using new incoming ML and pre-processing techniques makes the ETD process simple and efficient. Pamir et al. in [39] followed the same direction and proposed a hybrid ensemble model for electricity theft detection. The researchers worked on data pre-processing using KNNOR for data balancing. The feature reduction is done using the recursive feature elimination technique. For classification purposes, a bi-directional long short term memory (Bi-LSTM) classifier with three layers is used as the base model followed by a LogitBoost classifier. This proposed stacking approach results in improved detection performance when verified on a real-world 'SGCC' dataset. The output value obtained for precision, F1-score, and accuracy show 96.32%, 94.33%, and 89.45%, respectively.

3 | DESCRIPTION OF PROPOSED MODEL

In our proposed model, an ensemble AI technique is implemented for ETD in SG. The data is obtained from a utility company. The original data is prepared for model training using preprocessing and feature engineering steps. The entire dataset is split into training and testing sets. The training set is fed to base ML classifiers for training and prediction purposes. In the final step, the prediction from ML classifiers is used as features of a DL model for better classification results. The complete system, as shown in Figure 2, is divided into the following four steps.

1. The original data collected from the SG need to be preprocessed before feeding it into ML algorithm. It is because the original data have some missing values and outliers, and has a large variance. This may be due to hardware issues, noise in the communication medium, and users' different EC behavior. The missing values in the dataset decrease the model's performance. Therefore, they are replaced with mean values. To address the issue of outliers' handling, a Z-score capping technique is used in step-1. The large variation in the dataset reduces the model training capability. Therefore, the data is normalized using Min-Max scaling.
2. Step-2 addresses the issue of high dimensionality of the SGCC dataset. The large number of features cause time complexity issue and also reduce the model's performance having irrelavant features. This causes problems in ML model's data generalization. In our model the dataset is reduced using principal component analysis (PCA). This helps to increase storage efficiency and improve the model's performance by removing irrelevant features, and reduce storage cost and time complexity.
3. Step-3 shows the training data into the four ML models. These base models predict the output individually. These level-0 ML models include LGB, XGB, RF, and ET classifiers.
4. A multilayer perceptron (MLP) is used as a level-1 classifier in step 4, which obtains the output of level-0 models and predict the final output in the form of theft or normal electricity consumer.

3.1 | Dataset's information

The dataset used in this study is obtained from the real-time EC of Fujian city consumers connected to SGCC. This SGCC dataset, available as an MS Excel file, has a total of 42,372 consumers. There are mainly two types of consumers in this dataset, which are labeled as 0 and 1. Label-0 indicates a normal user while label-1 indicates a theft consumer. The consumers and their corresponding daily consumption are arranged as rows and columns in a table, which show the records and features of the dataset, respectively. Details of the dataset are organized in Table 2.

3.2 | Pre-processing

In ETD, we provide the users' ECP to the model and then use that for future cyber-physical theft attack (CPTA) predictions.

While the EC data obtained from the utility is un-scaled, imbalanced, and has missing values and outliers. Moreover, the information obtained from the original dataset, as shown in Table 2, cannot be used for accurate training of the model. Therefore, the data must be prepared using some ML techniques. The data also needs a proper or near to exact pattern for accurate detection. So, after the data is preprocessed, the

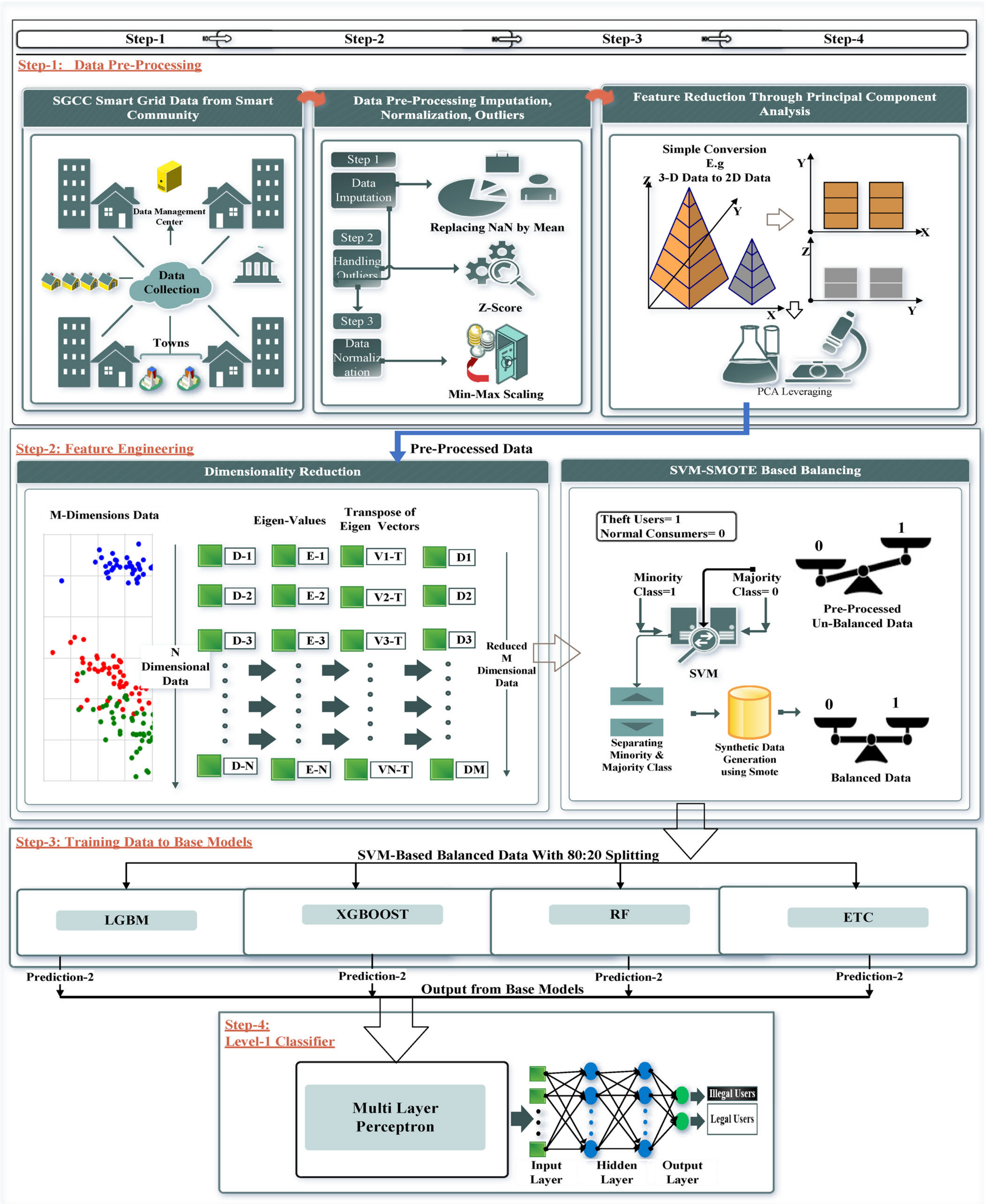

FIGURE 2 Proposed ETD stacked generalization model.

TABLE 2 SGCC original dataset's information.

Source of Data	Utility (SGCC)
Consumption duration	01/ 01/ 2014 to 31/ 10/ 2016
Consumers' category	Residential
Type of data	Daily real time consumption
Total number of consumers	42,372
Normal consumers	38,757
Theft consumers	3,615
Features	1,034
Records	42,372

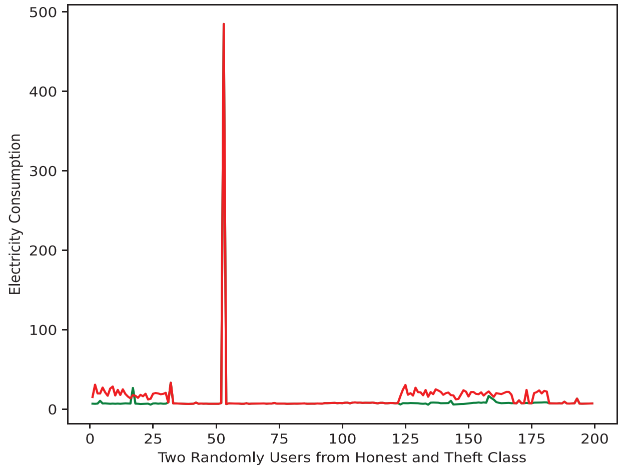

FIGURE 3 Electricity consumption pattern of two random consumers from SGCC dataset.

consumption patterns of the theft and normal users can be efficiently drawn. As shown in Figure 3, from a sample of the SGCC dataset, a normal user has a much smooth electricity usage pattern than a theft consumer that has large variations in the usage pattern. So the final pre-processed data can be used for model training and user behavior prediction. The pre-processing steps used in the proposed model are discussed below.

3.2.1 | Missing data imputation

The dataset obtained from the utility has a large number of missing values, denoted as not a number (NaN), which are different from each other. These NaN values may exist due to systematic, environmental, or random errors. The missing values cannot be neglected during preprocessing as they decrease the model's performance. Also, replacing them with zero will result in a loss of information. Many data science techniques are available for missing values imputation such as replacing NaN with mean, median, or mode. The median and mode cause a repetition of values in ECP, which again leads to performance performance deterioration. The imputation method given in

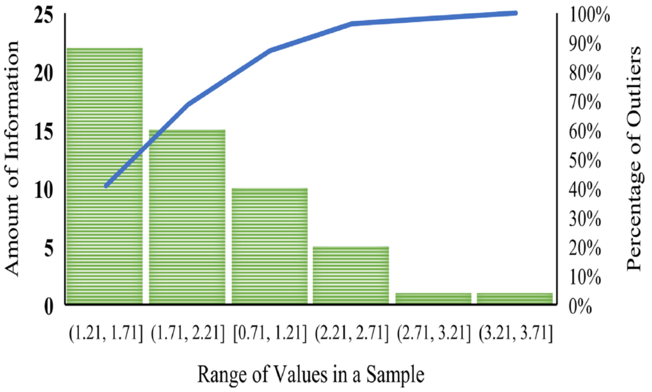

FIGURE 4 **FIGURE 4** Total contributions of outliers towards predictions.

[40] is used where NaN is imputed with a value, as given in the following equation. It has 3 main imputation conditions.

$$f(x) = \begin{cases} \dfrac{x_{(m,n)-1} + x_{(m,n)+1}}{2}, & \text{if } x_{m,n} = NaN, x_{(m,n)\pm1} \neq NaN \\ 0, & \text{if } x_{(m,n)\pm1} = NaN \\ x_{m,n} & \mathit{Otherwise} \end{cases}$$

In this equation, $x_{m,n}$ = the daily EC, $x_{m,n_{-1}}$ = the previous value, $x_{m,n_{+1}}$ = the next value to NaN.

3.2.2 | Handling outliers

In the ECP, we found some values to be too large or too small as compared to the normal values. These unexpected values (Outliers) deceive the model, which incurs a large execution time. Generally, the values below 10 percent and above 90 percent are treated as outliers.

As shown in Figure 4, the outliers in the dataset are less in number and show very little contribution towards model training. A novel z-score capping-based outliers' handling method [41], shown in algorithm-1, is applied to make the data more useful. The z-score outliers capping (ZSOC) technique works by first finding the z-score using Equation (1).

$$Z - score = Z = x_i - \frac{\mu}{\sigma^2}, \tag{1}$$

where $\sigma = \sqrt{\sum_{n=1}^{N} \frac{(x_n - \mu)^2}{n-1}}$ Z= standard score, x_i = random value of outliers, μ = the mean value,

sigma = standard deviation of row i. After calculating the Z-score, properly assign the lower and upper limits to the individual feature. The data point which is less than the lower limit or greater than the upper limit is replaced by the corresponding limit. The main advantage of using the capping technique is that it places the outlier at its respective extreme value instead of completely removing the entire row. This helps to retain the useful information in contrast to the present research [42, 43]

ALGORITHM 1 ZSOC working in cyber-physical electricity theft detection.

```
Input: Data X
Output: Data Y Start
Initialization: For i = 2,3,4,… N:
Original Dataset (X) is Selected
Find Z-score, Z:
for t ← 1 to TFind:
lower_limit = $\mu - 3 * \sigma$
the upper_limit= $\mu + 3\sigma$Upper limit imputationIf Z > upper_limit
Replace the Z with upper_limit Impute Lower limitIf Z < lower_limit
Replace the Z with upper_limitUpper limit imputationIf Z >
    upper_limit
Replace the Z with upper_limitImpute Lower limitIf Z < lower_limit
Replace the Z with upper_limitWhen no condition is satisfiedImpute:
    X_i= Z
Return ($Z_{m,n}$)
End
```

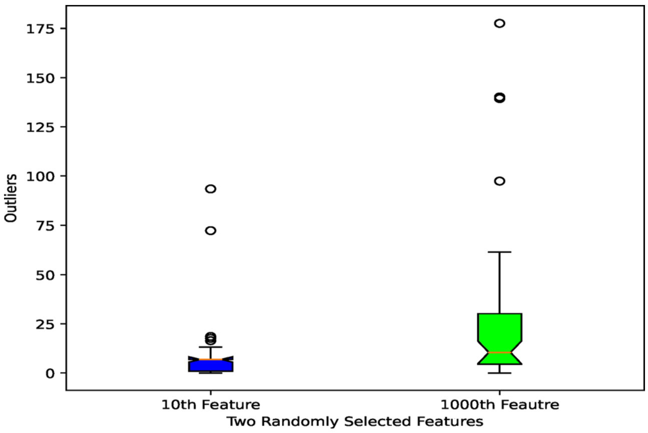

FIGURE 5 **FIGURE 5. Dots**show the presence of outliers.

where the outliers are entirely removed. Algorithm-1 presents the complete process.

3.2.3 | Unit based normalization

The Z-method was used to handle the outliers; however, the dataset still has large variations in the ECP of users, shown in Figure 5, for a sample taken from the SGCC dataset. These large variations degrade the output performance, as the ML and DL models are sensitive to the variation and quality of the dataset.

Min-max normalization from [31] is applied for scaling the data to the range [0, 1]. Min-max normalization has the mathematical form, shown in Equation (2).

$$f(x_{i,j}) = \frac{((x_{i,j}) - min(X))}{(max(X) - min(X))}. \tag{2}$$

Here min(X) and max(X) show the minimum and maximum EC of feature j in the data.

3.3 | Feature engineering

The data pre-processed in previous steps is fully prepared to be fed to the ML and DL models. But the ML and DL models face time complexity issues on a big and high-dimensional dataset. The feature engineering step is performed to reduce the size of the original dataset and also retain useful information. In the proposed system, PCA is used for feature reduction purpose. The PCA is used for reducing higher dimensional data into lower dimensions. This is obtained by forming linear relations among the features using mean and variance. The reduced features being obtained are called components that are independent of each other. This is due to the fact that PCA finds variance among the features and forms new components from the correlated features. The features which are more correlated are stored as individual components.

Similarly, the feature with the highest variance has more information and is stored as the first component. The feature with the second highest variance is taken as the second component, and so on. The overall PCA-based dimensionality reduction process [42] is given in Algorithm 2.

As the repeated and more related features are summed up as an individual component, it also reduces the over-fitting issue in the model. An additional arithmetic leveraging technique is also applied which results in improved performance in the proposed model. A set of 300 important features is extracted from the overall 1,034 features in the SGCC dataset. This helped in reducing the execution time. The main disadvantage of using PCA is that it cannot capture the minimum co-variance of the two classes and interpret the output features into such a uniform linear shape that it again leads to a small increase in simulation time. This issue is been tackled using an arithmetic leveraging technique, which also enhances the ECP separation.

3.4 | Data balancing

After missing value imputation, outlier removal, normalization, and feature engineering, the next step is to check for the class imbalance. In the SGCC dataset, the number of normal and abnormal users is not equally proportional. In the SGCC dataset, out of the total number of 42,372 users, the number of honest and illegitimate users are 38,757 and 3,615, respectively, as given in Table 2. The majority class (Normal=0) has more consumers than the minority class (Theft=1), as shown in Figure 6.

Due to this skewed behavior of the dataset, the machine learning model also shows a biased behavior towards the majority class and classifies the theft user (TU) as a normal consumer. To overcome the imbalance class issue, SVM-SMOTE is applied, which results in improved performance. The imbalanced dataset is shown in Figure 6 and the balanced data is given

ALGORITHM 2 Dimensionality reduction steps in principal component analysis.

```
Input: Data X
Output: Data Z
Initialization: Original Dataset, Y is Selected:
Chose Point of Interest
Find the Mean:
Mean = $\mu = \frac{x+x_{n-1}}{n}$
While$n \neq 0$ do where n = 2,3,4,… N$x_n$=values from dataset
x=point of interest
Find the Variance:
Id $\mu$ is Calculated Using$\sigma^2 = \frac{\sum_{n=1}^{N}(x_n-\mu)^2}{n}$$x_n$ =Each value in
    the datasetn = Number of values in the dataset$\mu$ = Mean of all the
    values
Find the Eigen Values ($\lambda$):
Determinant = Determinant [$A - \lambda * I$] A=Data Matrix, $\lambda$ =Eigen
    Value, $I$=Identity Matrix
If $\lambda$ is Determined
Compute the Eigen vector from Eigen Values:AX = $\lambda * X$A=
    N-dimensional dataX= N-Variables in dataset
If Eigen Vectors are Determined
Sort Vector in Descending order w.r.t lambda Values:
    $\lambda_n, \lambda_{n-1}, \lambda_{n-2}, \lambda_{n-3}, ... \lambda_{n-i}, ... \lambda_2, \lambda_1$
If done
Find the new Matrix W' from Eigen Vectors
Find Z from Y and W:
$Z = W * Y$
End
```

in Figure 7. Figure 7 shows a equal number of samples for malicious and normal users, after applying SVM-SMOTE technique. This helps to improve the ML model training by reducing the bias for majority class.

Figure 8 shows a visualization of how SVM-SMOTE works. The minority or theft class samples are less in number compared to normal users. So, SVM-SMOTE takes the available dataset and draw hyperplane between the two classes. The minority class samples are then considered and new samples are generated using K-nearest neighbor (KNN) algorithm. This makes equal the normal and theft class samples.

Different techniques are used in the present research such as RUS, random oversampling (ROS), and SMOTE. RUS reduces the records in the majority class to balance the dataset. But due to the reduction in the records, some important information is also lost, which decreases the model's performance. In contrast to RUS, ROS repeats random samples of the minority class and makes it equal to the majority class. Due to the repetition in the dataset, the over-fitting issue arises. To overcome the issues of RUS and ROS, SMOTE is used, which balances the dataset by generating synthetic data of the minority class. SMOTE also leads to oversampling of noise that causes high record dataset and causes time complexity and mis-classification issues. For proper data balancing and to overcome the issues in the above-used techniques, we propose a support vector machine minority oversampling technique (SVM-SMOTE) for better classification performance. SVM-SMOTE is a modified form of SMOTE, which is used for minority class oversampling. In this method, a hyperplane is drawn between the minority class and majority class, and synthetic data is generated on the minority side to obtain a balanced dataset. Also the new samples are generated closer to the boundary region because misclassification occurs due to samples closest to the boundary. So a clear boundary is obtained between the ECP of normal and malicious users, new synthetic samples are created and it becomes easy for model learning and future prediction. The SVM-SMOTE technique is used, shown in Figure 8, for balancing the SGCC dataset.

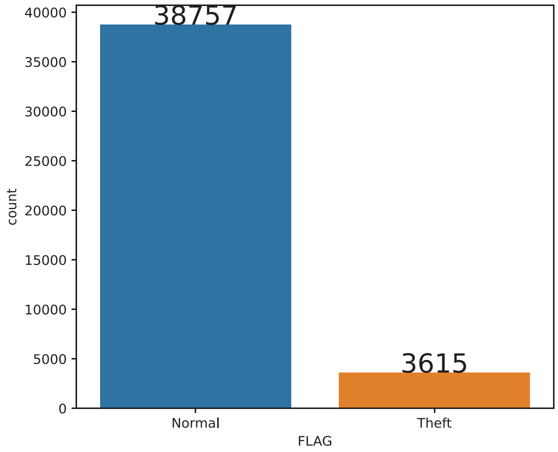

FIGURE 6 **FIGURE 6. Imbalanced** SGCC dataset.

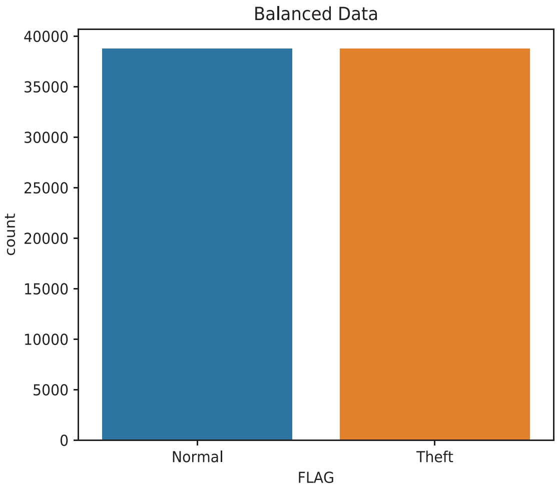

FIGURE 7 SVM-SMOTE based balanced dataset (SGCC).

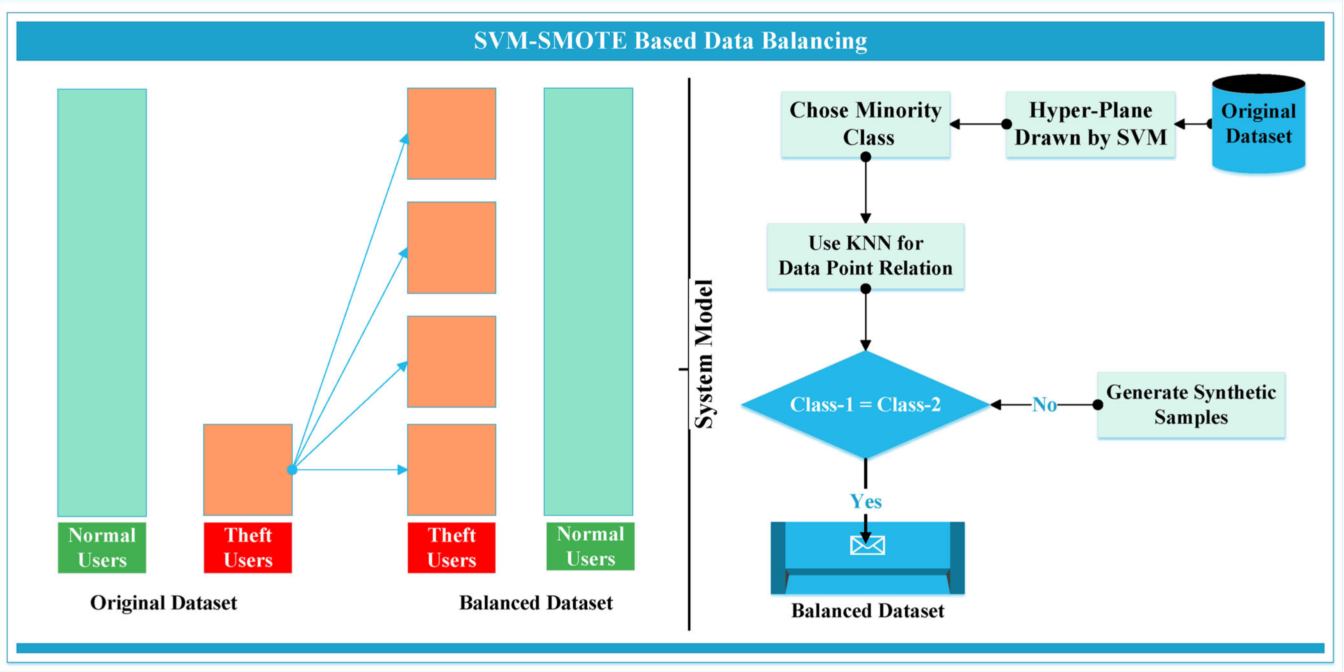

FIGURE 8 SVM-SMOTE system diagram.

3.5 | Model selection

In ML and DL, the time series and high dimensional datasets have enormous ECP and a single algorithm cannot learn and predict the accurate behavior. Four different ML models are considered in this work as weak learners to capture the ECP of all the customers for better generalization. These learners are LGB, RF, XGB, and ET. The details of the structure and process of all level-0 learners are provided as follows.

3.5.1 | Base learner-1

Light gradient boosting (LGB) released by Microsoft in 2017 [44]. LGB is a modified form of gradient boosting tree algorithm with leaf-wise splitting for higher accuracy. Due to the leaf-wise splitting structure, LGB is useful for complex modeling like time series classification, regression, ranking etc.

3.5.2 | Base learner-2

Random forest (RF) is an ensemble ML algorithm, which is used for classification and regression. The algorithm is simple in structure with many DTs. RF is the best tool used for multivariable datasets. It is a widely used algorithm as it can produce good results without hyper-parameter optimization. The basic mechanism of this model is that it uses the bootstrapping phenomenon. Where the original dataset is randomly divided into subsets with replacements [45]. These bootstraps are then used for DT where each tree makes a prediction. Based on these predictions, a voting mechanism is implemented. This gives rise to the final prediction of the RF model. RF can handle large datasets, reduce the variance and over-fitting, and show a higher accuracy as compared to the DT classifier. The Gini index is a statistical term used to predict the outcome probability of a random forest. Mathematically, Gini index can be found using Equation (3) [46].

$$Gini\_index = 1 - \sum_{i=1}^{c} (p_i)^2 \tag{3}$$

Here, c = Number of classes

p_i = Relative frequency of the given class outcome

3.5.3 | Base learner-3

XGB or regularized gradient boosting is a sequential tree-based algorithm that focuses on computational speed and model performance. This algorithm basically works on the Taylor series function to find the loss function [47]. The model combines the weak learners sequentially to improve their learning. A new regularization term is included to prune the extra leaf and avoid overfitting. The algorithm can be used for both regression and classification tasks and has been designed to work with large and complicated datasets.

3.5.4 | Base learner-4

Extra tree classifier (ETC), also named extremely randomized tree, is a DT-based bagging technique. It uses training data and creates a large number of random un-pruned trees. In the final step, ETC reduces the model training by collecting a random DT for the best split [48]. Due to the random pruning phenomenon and in the absence of an optimum splitting step, ETC has a very short execution time and is this applied in this model.

3.5.5 | Stacking model

The main purpose of building a stacking ML model is to obtain better classification results, specifically theft detection in SG.

The model produces more accurate results than the individual classifier. The stacked generalization combines the learning ability of multiple algorithms for achieving optimum accuracy in terms of classification [49]. The proposed system combines the strength of all four level-0 classifiers for a reduction in variance, bias, overfitting, and execution time. The model deals with big data, and makes accurate predictions. In the stacking model, the training dataset is fed to the base learners with k-fold cross-validation. The level-0 classifiers learn to make predictions on the out-of-fold dataset. In the next step, the predictions from all the base learners are used as features of the level-1 classifier or meta-classifier. The meta-classifier learns the predictions of level-0 learners and predicts the output class. The complete stacking process is shown in Algorithm-3.

ALGORITHM 3 Proposed stacking generalization technique for theft detection.

```
Input: Data X
Output: Data Z
Original Data=X= $\sum X^{N}_{(i,j,k,l,m,n,o=1)}$
Select the Original Dataset
Split the data:
Training set=X= $\sum X1^{N}_{(i,j,k,l,m=1)}$
Testing set=Y= $\sum Y1^{N}_{(n,o=1)}$
Level-0 classifier (C1):If $n \neq 0$ andfor t ← 1 to T= X
Learn base classifier C1 on X1= $\sum X1^{N}_{(i,j,k,l=1)}$
Predict C1 on Validation set V1= $\sum Y1^{N}_{(m=1)}$
0utput prediction=P1
Level-0 classifier (C2): C1 is Calculated
Learn base classifier C2 on X2= $\sum X2^{N}_{(i,j,k,m=1)}$
Predict C2 on the Validation set V2= $\sum Y2^{N}_{(l=1)}$
Output prediction=P2
Find C3, If C2 is Determined FindC3,
IfC1, C2 is Determined
Level-0 classifier (C3):
Similarly, learn the base classifiers C4
Make predictions P4
After All Level-0 are Predicted
Meta classifier (M1)
Learn Meta Classifier M1 on X=[P1, P2, P3, P4]
Predict M1 on Y
End
```

3.6 | MLP mathematical modeling

An MLP is a useful tool for non-linear data classification. It has three main layers, input layer, hidden layer and output layer, and each of these layers serves a specific purpose in the network's overall architecture.

1. An input layer the number of input layers depends on input features its function is to receive the input data and pass it on to the next layer in the network. The number of nodes in the input layer is equal to the number of features in the input data. For example, in our case, the number of features used is 300, so 300 input layers are used.
2. The hidden layers are used for weight updation. The hidden layer is the layer between the input and output layers. Each node in the hidden layer takes information from nodes in the layer below and creates an output that is sent to the layer above. The number of nodes in each of the hidden layers of a neural network might vary.
3. The neural network's last layer, known as the output layer, is responsible for producing the final output prediction based on the input data. The type of problem being solved determines how many nodes are present in the output layer. In our binary classification issue, for instance, the output layer would contain two nodes: one representing the likelihood of belonging to the honest class and the other representing the probability of belonging to the theft class.

The input layers provide a scaled signal to the hidden layers. The weights are real numbers multiplied by the input signals. The hidden layers give the weighted sum of the given information [50].

$$y_o = \sum_{i=1}^{n} w_i x_i + b. \tag{4}$$

The above-obtained information is still in linear form. The activation function given below is used to obtain information about non-linear data.

$$f(x) = \frac{1}{1 + e^{-x}}. \tag{5}$$

Then the information obtained from hidden layers is found using the equation given below.

$$y_o = f(x)(\sum_{i=1}^{n} w_i x_i + b), \tag{6}$$

where y_o is the output, w_i is the weight value, x_i is input data, b is the bias factor and *f(x)* is the activation function. The number of neurons determines the hidden layers in the network.

If the number of neurons is kept very then small it will lead to model under-fitting while a large number of neurons lead to an over-fitting issue in the model prediction. A default sigmoid activation function is used in the network for non-linear data modeling. The limit of sigmoid activation is between 0 to one, with 0 for negative values and 1 for positive values. The overall equation used for MLP is given below.

$$y_o = f[WO_{mn}(\sum_{i=1}^{n} WI_{ij}x_i + b_1) + b_2], \tag{7}$$

where WI_{ij} is the weight of input layer, WO_{mn} is the weight of output layer, b_1 is the input bias factor and b_2 is the bias in output layer.

4 | PERFORMANCE METRICS

For classification problems, various performance parameters are used to evaluate the final output and performance of the model like confusion matrix, F1-score, Area Under the Curve, Precision, Recall, Receiver Operating Curve, and Accuracy. These parameters are helpful in checking the overall performance of a model. The parameters are explained with respective mathematical forms in the following paragraphs [51].

4.1 | Confusion matrix

In ML, a confusion matrix is used to measure classification performance. It is an N * N matrix. Here N is the number of classes in a given dataset. The matrix has 2 dimensions with actual class and predicted class. In our SGCC data set, we have binary (2) class classifications normal (0) and malicious (1). So the confusion matrix is 2*2 matrix [46]. And has the following 4-types of outputs.

(a) True Positive (TP) = True positive is the actual positive class (1) value which is predicted positive (1) by the classifier.
(b) True Negative (TN) = True negative is negative class (0) data points which is predicted as negative (0) by the model.
(c) False Positive (FP) = False positive is the negative class (0) values and the model classify it as positive class (1).
(d) False Negative (FN) = And false negative shows the positive class (1) values which is predicted as negative class (0) by the ML model.

4.2 | Accuracy

In ML accuracy is used to measure the overall performance of a model on a given dataset. It is used to measure how much data is classified correctly. Considering the confusion it is the number of correctly classified data points divided by all the data points predicted by a given ML model. Mathematically, accuracy is calculated using Equation (8) [52].

$$Accuracy = \frac{TP * TN}{TP * TN * FP * FN}. \tag{8}$$

After data scaling, there are still some outliers in the dataset, which disturb the model's performance and learning time. The Z-score-based capping technique is used in this paper to properly address the outliers. The main advantage of using the capping technique is that it places the outliers at their respective extreme value instead of completely removing the entire row.

4.3 | Precision

Precision is the portion of positive data points that are correctly classified. It is the positive class values, from all the predicted values, that the model classified as positive. Precision is sometimes referred to as specificity. The following mathematical formula, given in Equation (9), used to find the precision [50].

$$Precision(P) = \frac{TP}{TP * FP}. \tag{9}$$

4.4 | Recall

The recall represents the number of positively classified data points. It is the portion of actual positive values that the model classifies as positive and negative [46]. The recall is also called sensitivity and has the following formula (10):

$$Recall\,(R) = \frac{TP}{TP * FN}. \tag{10}$$

4.5 | F1-score

In the classification cases, the main aim is to obtain the best value for precision and recall. F1-score is the measure used to find the best classification values in terms of precision and recall [50]. Mathematically, it is the harmonic mean of precision and recall, and is given in Equation (11).

$$F1 - score = 2 * \frac{Precision * Recall}{Precision + Recall}. \tag{11}$$

4.6 | Area under the curve

AUC is the total area covered by the ROC curve or the total points lying under the ROC curve. The threshold with maximum ROC is called the AUC value of that model [46].

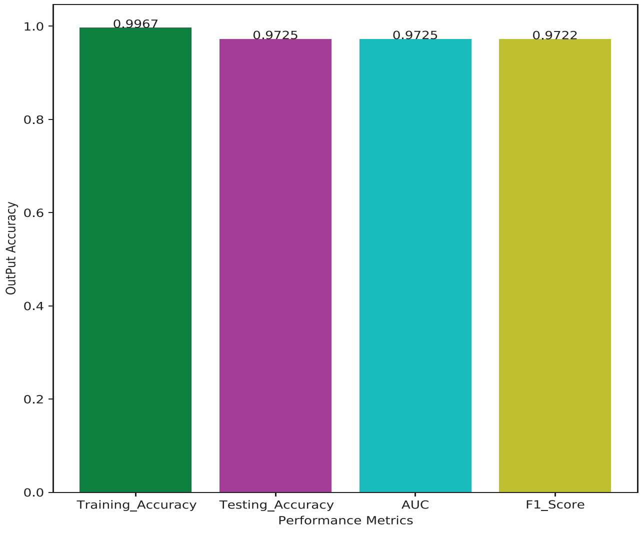

FIGURE 9 Proposed model performance on SGCC dataset.

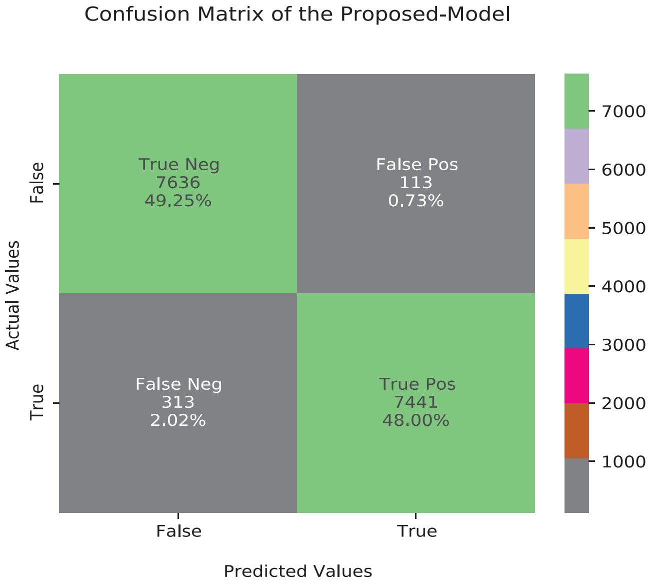

FIGURE 10 Proposed model confusion matrix on SGCC dataset.

4.7 | Receiver operating characteristics

The ROC curve shows the TP predicted values at different thresholds w.r.t FP points. ROC is important to measure when dealing with imbalanced datasets. More specifically, it is the graph of true positive rate (TPR) w.r.t false positive rate (FPR) [52, 53]. Equation (12) is used to find the TPR:

$$TPR = \frac{TP}{TP + FN}. \tag{12}$$

While to find the FPR, Equation (13) is used:

$$FPR = \frac{FP}{FP + TN}. \tag{13}$$

5 | SIMULATION SETUP

In this section, we set the prepared and reduced dataset into training and testing subsets. The processed data obtained from the above three steps are split into 80:20 for model training and testing. The final result is shown in the next section.

6 | RESULTS' DISCUSSION AND EVALUATION

The results obtained are evaluated in the form of important performance metrics required for classification purposes. The experimental results obtained after the model simulations are discussed as follows. Accuracy is a general classification term that may not be a good metric for classification. Therefore, a combination of different performance metrics is used for good classification. Figure 9 shows the training and testing accuracy, F1-score, and AUC of the proposed model.

The training accuracy is 0.9967 or 99.67% which shows that our model well trained and have good learning capacity. The model showed 0.9725 or 97.25% accuracy that proves that the proposed model is useful to be used for prediction purposes. The AUC and F1-score, which are other performance parameters, also confirm the model testing performance.

The confusion matrix obtained in Figure 10 shows the final prediction of the proposed model in terms of TP, TN, FP, and FN. As seen from the figure that the proposed model has a high detection rate for normal and theft class prediction. The values of FP and FN represent wrongly classified users. A reduction in these parameters is obtained, with the misclassification of FP and FN to 0.54% and 1.69%, respectively. This achieves our proposed objective in terms of very low FPR and FNR. The output performance of different models on pre-processed data is given in Table 3.

The precision and recall values are usually calculated in a single relationship. This combined precision-recall curve (PRC) is obtained for different threshold values. The high precision shows a low FP value and a high recall represents a lower value of FN. The PRC curve, shown in Figure 11, obtained using the proposed model shows both precision and recall have a high value of 97.1%.

In ML accuracy shows how much of the data points are correctly classified out of the total predicted points. This clarifies how many of the users are predicted malicious and how many are predicted as normal using the ML model. Figure 12 shows a bar chart with accuracy values of all base models and the proposed model. The proposed model achieved a high accuracy of 97.6% as compared to level-0 models. The values are given in Table 4.

The ROC plots the TPR and the FPR at different thresholds. The high value of ROC shows the positive class prediction ability. Figure 13 shows the ROC value of the proposed model to be 96.7%.

TABLE 3 **TABLE 3. Output** performance of different models on pre-processed data.

Model + Applied technique (s)	Accuracy in %	FPR in %	FNR in %	Time (s)
PCA + SMOTE	95.95	1.31	2.73	1320
PCA + SVM-SMOTE	96.33	1.09	2.57	1540
Z-score-capping + PCA + SVM-SMOTE	96.27	1.12	2.61	1220
Z-score-capping + PCA(features=200) + arithmetic-leveraging +SVM-SMOTE	97.3	0.60	2.04	1850
Z-score-capping + PCA(features=400) + arithmetic-leveraging +SVM-SMOTE	97.29	0.62	2.09	2520
Z-score-capping + PCA(features=300) + arithmetic-leveraging + SVM-SMOTE	97.69	0.60	1.82	2070

TABLE 4 Models' performance on different data splitting.

Data splitting model	Train/test size = 80:20				Train/test size = 75:25				Train/Test Size = 70: 30			
	Tr. Acc	Tes. Acc	AUC	F1-score	Tr. Acc	Tes. Acc	AUC	F1-score	Tr. Acc	Tes. Acc	AUC	F1-score
LGBM	0.9478	0.9329	0.9330	0.9315	0.9487	0.9293	0.9292	0.9275	0.9480	0.9309	0.9308	0.9291
RF	0.9940	0.9493	0.9494	0.9488	0.9943	0.9487	0.9487	0.9483	0.9937	0.9444	0.9443	0.9436
XGBoost	0.8329	0.8239	0.8240	0.8168	0.8353	0.8240	0.8239	0.8197	0.8334	0.8304	0.8306	0.8242
ET	0.9992	0.8252	0.8252	0.8319	0.9984	0.8044	0.8046	0.8131	0.9993	0.8069	0.8070	0.8151
Proposed model	99.78	0.9769	0.9769	0.9766	0.9974	0.9754	0.9753	0.9749	0.9982	0.9743	0.9743	0.9739

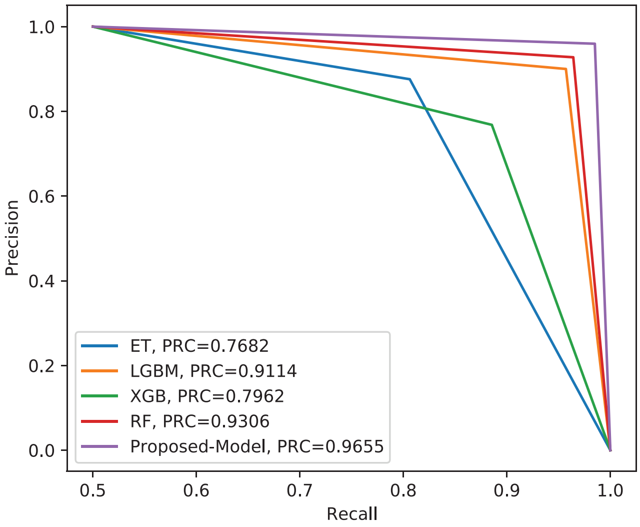

FIGURE 11 Precision-recall curve of the base models and proposed model.

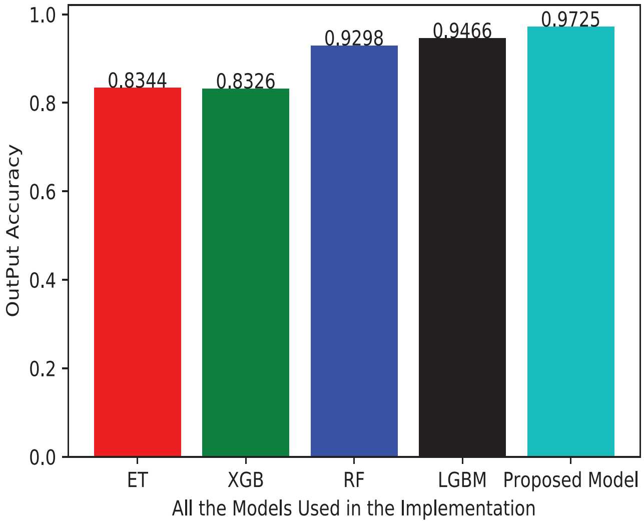

FIGURE 12 Accuracy comparison of level-0 and the proposed model.

In ML, AUC is a two-dimensional curve with TP on the y-axis and FP on the x-axis. The AUC aggregates the TP and FP values on all given thresholds. A high value of AUC suggests a higher prediction of a positive class, which is electricity theft in our case.

Figure 14 presents the AUC, F1-score, and accuracy values of all the level-0 models and the proposed model. The results given in Table 5 also shows a higher performance of the proposed model as compared to the base models.

TABLE 5 F1-score, AUC, accuracy of the base models, and proposed model.

Model	F1-score	AUC	Accuracy
XGBoost	0.8168	0.8240	0.8239
ET	0.8319	0.8252	0.8252
LGBM	0.9315	0.9330	0.9329
RF	0.9488	0.9494	0.9493
Proposed model	0.9766	0.9769	0.9769

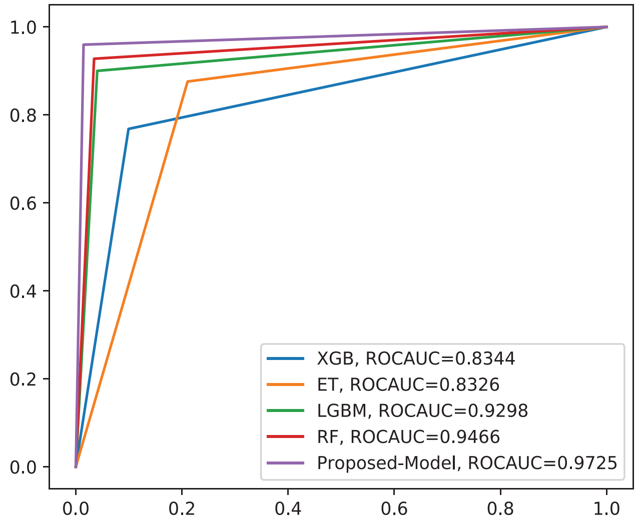

FIGURE 13 Comparison of base models' ROC with the proposed model's.

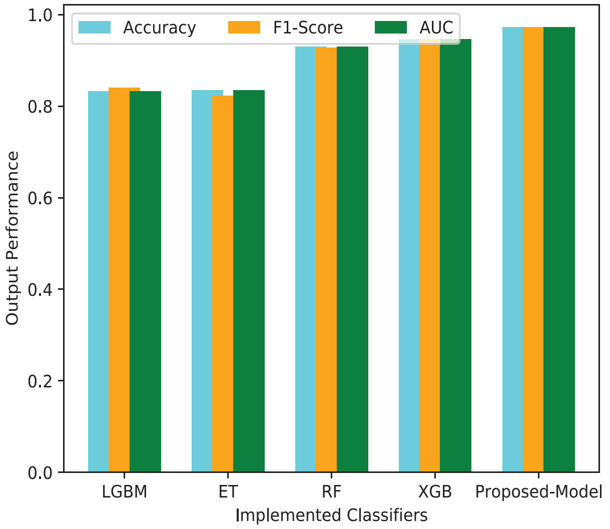

FIGURE 14 Comparison of AUC, F1-score and accuracy of base models and the proposed model.

7 | CONCLUSION

In this paper, an ML-based stacked generalization technique is proposed to overcome the CPTA issue and cut down on the anomalous consumption of electricity happening in the smart cities. The overall system is divided into four modules with specific functions. The data obtained from the utility needs some pre-processing before being used for model training. The first module addresses these issues with novel techniques in order to process the data without losing important information. The NaN is imputed using the mean imputation method for making a complete ECP. The data is normalized with the min-max scaling technique to bring the data into a proper range. A z-score capping technique is applied for efficient handling of the outliers in the dataset. In the second module, a leveraging-PCA-based technique is applied for important feature extraction and data reduction purposes. We implement the SVM-SMOTE technique for optimal balancing of the theft and normal class data obtained from the PCA technique. The benchmark classifiers are implemented in module 3 of the proposed model. The dataset is split into 80:20 training and testing ratio after balancing, and fed to the four base classifiers. These base classifiers are trained and predicted on 80% (training set) of the dataset, and the predictions are obtained for each classifier. The final classification is performed in module 4, with input from four different ML models and a meta-level DL model. The prediction of level-0 classifiers is fed to the level-1 model to capture the ECP information from all the base classifiers. The final prediction obtained from the level-1 model shows an enhanced performance in terms of classification. The results obtained show that our proposed model outperformed other benchmark ML models. The proposed model achieves a high accuracy of 97.69%. In addition, a very low value of FPR and FNR is obtained with 0.72%, 2.05%, respectively, which show the reduction in inspection cost and energy theft. The results obtained make the proposed model useful to be used in industrial applications for theft detection and NTL reduction purposes.

8 | FUTURE WORK

The proposed system model used here works in offline mode. In the future, the concern will be to transform the model into the online mode with hyper-parameter tuning, using a heuristic algorithm, in order to address the system execution time and improve the detection accuracy in the smart grids.

AUTHOR CONTRIBUTIONS

Arshid Ali: Writing – original, conceptualization, methodology, software. **Laiq Khan**: Supervision, writing – original, validation, software. **Nadeem Javaid**: Supervision, writing – review & editing, validation. **Safdar Hussain Bouk**: Visualization, investigation. **Abdulaziz Aldegheishem**: Project supervision, funding, validation. **Nabil Alrajeh**: Writing – review & editing, software.

ACKNOWLEDGEMENTS

The authors would like to acknowledge the support of Researchers Supporting Project Number RSP2023R295, King Saud University, Riyadh, Saudi Arabia.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interests.

DATA AVAILABILITY STATEMENT

Not applicable.

ORCID

Arshid Ali https://orcid.org/0000-0002-9657-1494
Nadeem Javaid https://orcid.org/0000-0003-3777-8249

REFERENCES

1. U.S. Energy Information Administration (EIA). https://www.eia.gov/tools/faqs/faq.php?id= 427t=3. Accessed December 2022.
2. Ullah, A., Javaid, N., Asif, M., Javed, M.U., Yahaya, A.S.: Alexnet, adaboost and artificial bee colony based hybrid model for electricity theft detection in smart grids. IEEE Access 10, 18681–18694 (2022)
3. Massaferro, P., Di Martino, J.M., Fernández, A.: Fraud detection on power grids while transitioning to smart meters by leveraging multi-resolution consumption data. IEEE Trans. Smart Grid 13(3), 2381–2389 (2022)
4. Shah, A.L., Mesbah, W., Al-Awami, A.T.: An algorithm for accurate detection and correction of technical and nontechnical losses using smart metering. IEEE Trans. Instrum. Meas. 69(11), 8809–8820 (2020)
5. Lepolesa, L.J., Achari, S., Cheng, L.: Electricity theft detection in smart grids based on deep neural network. IEEE Access 10, 39638–39655 (2022)
6. Javaid, N.: A PLSTM, alexNet and ESNN based ensemble learning model for detecting electricity theft in smart grids. IEEE Access 9, 162935–162950 (2021)
7. Saleem, M.U., Usman, M.R., Usman, M.A., Politis, C.: Design, deployment and performance evaluation of an IoT based smart energy management system for demand side management in smart grid. IEEE Access 10, 15261–15278 (2022)
8. Chamra, A., Harmanani, H.: A smart green house control and management system using IoT. In: 17th International Conference on Information Technology-New Generations (ITNG 2020), pp. 641–646. Springer International Publishing, Cham (2020)
9. Buzau, M.M., Tejedor-Aguilera, J., Cruz-Romero, P., Gómez-Expósito, A.: Detection of non-technical losses using smart meter data and supervised learning. IEEE Trans. Smart Grid 10(3), 2661–2670 (2018)
10. Mujeeb, S., Javaid, N., Ahmed, A., Gulfam, S.M., Qasim, U., Shafiq, M., Choi, J.-G.: Electricity theft detection with automatic labeling and enhanced RUSBoost classification using differential evolution and Jaya algorithm. IEEE Access 9, 128521–128539 (2021)
11. Hamad, A.A., Abdulridha, M.M., Kadhim, N.M., Pushparaj, S., Meenakshi, R., Ibrahim, A.M.: Learning methods of business intelligence and group related diagnostics on patient management by using artificial dynamic system. J. Nanomater. 2022, 1–8 (2022)
12. Yan, Z., Wen, H.: Electricity theft detection base on extreme gradient boosting in AMI. IEEE Trans. Instrum. Meas. 70, 1–9 (2021)
13. Rani, S., Babbar, H., Srivastava, G., Gadekallu, T.R., Dhiman, G.: Security framework for internet of things based software defined networks using blockchain. IEEE Internet Things J. 10(7), 6074–6081 (2023)
14. Arif, A., Alghamdi, T.A., Ali Khan, Z., Javaid, N.: Towards efficient energy utilization using big data analytics in smart cities for electricity theft detection. Big Data Res. 27, 100285 (2022)
15. Li, D., Deng, L., Gupta, B.B., Wang, H., Choi, C.: A novel CNN based security guaranteed image watermarking generation scenario for smart city applications. Inf. Sci. 479, 432–447 (2019)
16. Al-Rahawe, B.A., Hamad, A.A., Al-Zuhairy, M.H., Khalaf, H.H., Abebaw, S.: The commitment of Nineveh governorate residents to the precautionary measures against global 2019 pandemic and dermatological affection of precautions. Appl. Bionics Biomech. 2021, 1526931 (2021)
17. Yan, Z., Wen, H.: Performance analysis of electricity theft detection for the smart grid: An overview. IEEE Trans. Instrum. Meas. 71, 1–28 (2021)
18. Pasdar, A., Mirzakuchaki, S.: A solution to remote detecting of illegal electricity usage based on smart metering. In: 2007 2nd International Workshop on Soft Computing Applications, pp. 163–167. IEEE, Piscataway (2007)
19. Ali, S.S., Maroof, M., Hanif, S.: Smart energy meters for energy conservation and minimizing errors. In: 2010 Joint International Conference on Power Electronics, Drives and Energy Systems and 2010 Power India, pp. 1–7. IEEE, Piscataway (2010)
20. Zheng, D., Wang, S.: Research on measuring equipment of single-phase electricity-stealing with long-distance monitoring function. In: 2009 Asia-Pacific Power and Energy Engineering Conference, pp. 1–4. IEEE, Piscataway (2009)
21. Astronomo, J., Dayrit, M.D., Edjic, C., Regidor, E.R.T.: Development of electricity theft detector with GSM module and alarm system. In: 2020 IEEE 12th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment, and Management (HNICEM), pp. 1–5. IEEE, Piscataway (2020)
22. Somefun, T.E., Awosope, C.O.A., Chiagoro, A.: Smart prepaid energy metering system to detect energy theft with facility for real time monitoring. Int. J. Electr. Comput. Eng. 9(5), 4184 (2019)
23. Saritha, G., Sowmyashree, M.S., Thejaswini, S., Surekha, R.G.: Wireless power theft monitoring and controlling unit for substation. IOSR J. Electron. Commun. Eng. 9(1), 10–14 (2014)
24. Mir, S.H., Ashruf, S., Bhat, Y., Beigh, N.: Review on smart electric metering system based on GSM/IOT. Asian J. Electr. Sci. 8(1), 1–6 (2019)
25. Fovino, I.N., Carcano, A., De Lacheze Murel, T., Trombetta, A., Masera, M.: Modbus/DNP3 state-based intrusion detection system. In: 2010 24th IEEE International Conference on Advanced Information Networking and Applications, pp. 729–736. IEEE, Piscataway (2010)
26. Bandim, C.J., Alves, J.E.R., Pinto, A.V., Souza, F.C., Loureiro, M.R.B., Magalhaes, C.A., Galvez-Durand, F.: Identification of energy theft and tampered meters using a central observer meter: a mathematical approach. In: 2003 IEEE PES Transmission and Distribution Conference and Exposition (IEEE Cat. No. 03CH37495), vol. 1, pp. 163–168. IEEE, Piscataway (2003)
27. McLaughlin, S., Holbert, B., Fawaz, A., Berthier, R., Zonouz, S.: A multi-sensor energy theft detection framework for advanced metering infrastructures. IEEE J. Sel. Areas Commun. 31(7), 1319–1330 (2013)
28. Xia, X., Xiao, Y., Liang, W., Zheng, M.: GTHI: A heuristic algorithm to detect malicious users in smart grids. IEEE Trans. Network Sci. Eng. 7(2), 805–816 (2018)
29. Cárdenas, A.A., Amin, S., Schwartz, G., Dong, R., Sastry, S.: A game theory model for electricity theft detection and privacy-aware control in AMI systems. In: 2012 50th Annual Allerton Conference on Communication, Control, and Computing (Allerton), pp. 1830–1837. IEEE, Piscataway (2012)
30. Gao, Y., Foggo, B., Yu, N.: A physically inspired data-driven model for electricity theft detection with smart meter data. IEEE Trans. Ind. Inf. 15(9), 5076–5088 (2019)
31. Jindal, A., Dua, A., Kaur, K., Singh, M., Kumar, N., Mishra, S.: Decision tree and SVM-based data analytics for theft detection in smart grid. IEEE Trans. Ind. Inf. 12(3), 1005–1016 (2016)
32. Cvitić, I., Peraković, D., Periša, M., Gupta, B.: Ensemble machine learning approach for classification of IoT devices in smart home. Int. J. Mach. Learn. Cybern. 12(11), 3179–3202 (2021)
33. Yan, Z., Wen, H.: Electricity theft detection base on extreme gradient boosting in AMI. IEEE Trans. Instrum. Meas. 70, 1–9 (2021)
34. Punmiya, R., Choe, S.: Energy theft detection using gradient boosting theft detector with feature engineering-based preprocessing. IEEE Trans. Smart Grid 10(2), 2326–2329 (2019)
35. Panthi, M.: Anomaly detection in smart grids using machine learning techniques. In: 2020 First International Conference on Power, Control and Computing Technologies (ICPC2T), pp. 220–222. IEEE, Piscataway (2020)
36. Jokar, P., Arianpoo, N., Leung, V.C.M.: Electricity theft detection in AMI using customers' consumption patterns. IEEE Trans. Smart Grid 7(1), 216–226 (2015)
37. Ayub, N., Aurangzeb, K., Awais, M., Ali, U.: Electricity theft detection using CNN-GRU and manta ray foraging optimization algorithm. In: 2020 IEEE 23rd International Multitopic Conference (INMIC), pp. 1–6. IEEE, Piscataway (2020)
38. Ghori, K.M., Abbasi, R.A., Awais, M., Imran, M., Ullah, A., Szathmary, L.: Performance analysis of different types of machine learning classifiers for non-technical loss detection. IEEE Access 8, 16033–16048 (2019)
39. Javaid, N., Almogren, A., Adil, M., Javed, M.U., Zuair, M.: RFE based feature selection and KNNOR based data balancing for electricity theft detection using BiLSTM-LogitBoost stacking ensemble model. IEEE Access 10, 112948–112963 (2022)
40. Javaid, N., Javaid, S., Asif, M., Javed, M.U., Yahaya, A.S., Aslam, S.: Synthetic theft attacks and long short term memory-based preprocessing for

electricity theft detection using gated recurrent unit. Energies 15(8), 2778 (2022)

41. Analytics Vidhya: Dealing with outliers using the z-score method. https://www.analyticsvidhya.com/blog/2022/08/. Accessed December 2022.
42. Shehzad, F., Javaid, N., Almogren, A., Ahmed, A., Gulfam, S.M., Radwan, A.: A robust hybrid deep learning model for detection of non-technical losses to secure smart grids. IEEE Access 9, 128663–128678 (2021)
43. Ünal, F., Almalaq, A., Ekici, S., Glauner, P.: Big data-driven detection of false data injection attacks in smart meters. IEEE Access 9, 144313–144326 (2021)
44. Kanna, P.R., Sindhanaiselvan, K., Vijaymeena, M.K.: A defensive mechanism based on PCA to defend denial of-service attack. Int. J. Sec. Appl. 11(1), 71–82 (2017)
45. Tao, P., Shen, H., Zhang, Y., Ren, P., Zhao, J., Jia, Y.: Status forecast and fault classification of smart meters using LightGBM algorithm improved by random forest. Wireless Commun. Mobile Comput. 2022, 3846637 (2022)
46. Lin, G., Feng, X., Guo, W., Cui, X., Liu, S., Jin, W., Lin, Z., Ding, Y.: Electricity theft detection based on stacked autoencoder and the undersampling and resampling based random forest algorithm. IEEE Access 9, 124044–124058 (2021)
47. Daniya, T., Geetha, M., Kumar, K.S.: Classification and regression trees with Gini index. Adv. Math. Sci. J. 9(10), 8237–8247 (2020)
48. Patnaik, B., Mishra, M., Bansal, R.C., Jena, R.K.: MODWT-XGBoost based smart energy solution for fault detection and classification in a smart microgrid. Appl. Energy 285, 116457 (2021)
49. Acosta, M.R.C., Ahmed, S., Garcia, C.E., Koo, I.: Extremely randomized trees-based scheme for stealthy cyber-attack detection in smart grid networks. IEEE Access 8, 19921–19933 (2020)
50. Ouyang, Z., Sun, X., Chen, J., Yue, D., Zhang, T.: Multi-view stacking ensemble for power consumption anomaly detection in the context of industrial internet of things. IEEE Access 6, 9623–9631 (2018)
51. Mosavi, M.R., Khishe, M., Naseri, M.J., Parvizi, G.R., Mehdi, A.Y.A.T.: Multi-layer perceptron neural network utilizing adaptive best-mass gravitational search algorithm to classify sonar dataset. Arch. Acoust. 44(1), 137–151 (2019)
52. Javaid, N., Qasim, U., Yahaya, A.S., Alkhammash, E.H., Hadjouni, M.: Non-technical losses detection using autoencoder and bidirectional gated recurrent unit to secure smart grids. IEEE Access 10, 56863–56875 (2022)
53. Khattak, A., Bukhsh, R., Aslam, S., Yafoz, A., Alghushairy, O., Alsini, R.: A hybrid deep learning-based model for detection of electricity losses using big data in power systems. Sustainability 14(20), 13627 (2022)

How to cite this article: Ali, A., Khan, L., Javaid, N., Bouk, S.H., Aldegheishem, A., Alrajeh, N.: Mitigating anomalous electricity consumption in smart cities using an AI-based stacked-generalization technique. IET Renew. Power Gener. 1–18 (2023). https://doi.org/10.1049/rpg2.12785